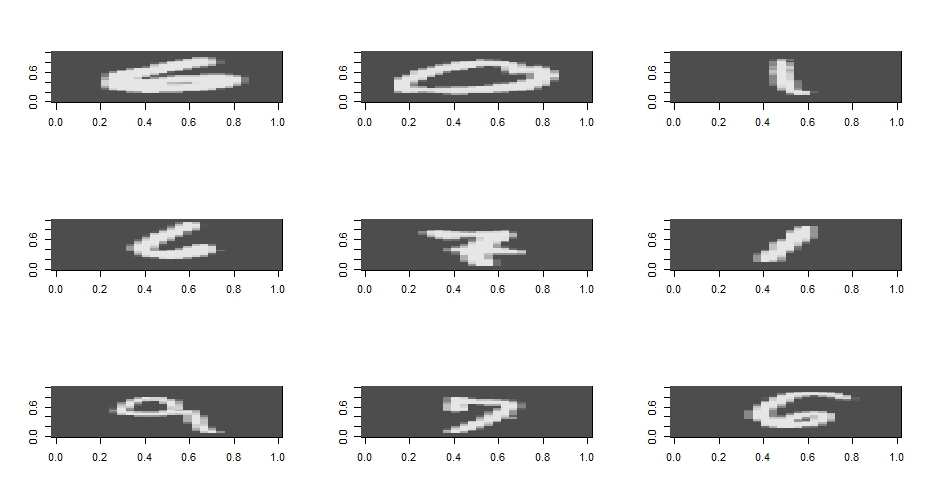
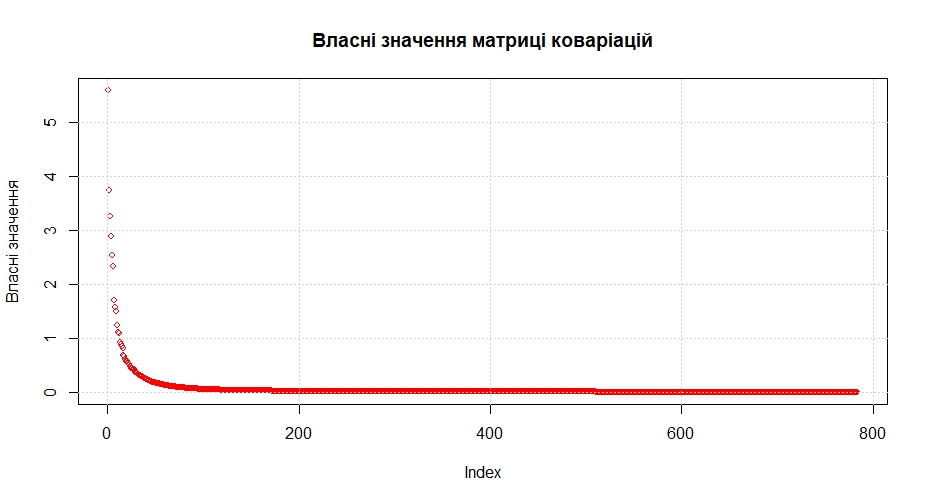
Лабораторна робота 4

Методи та засоби кластерного аналізу

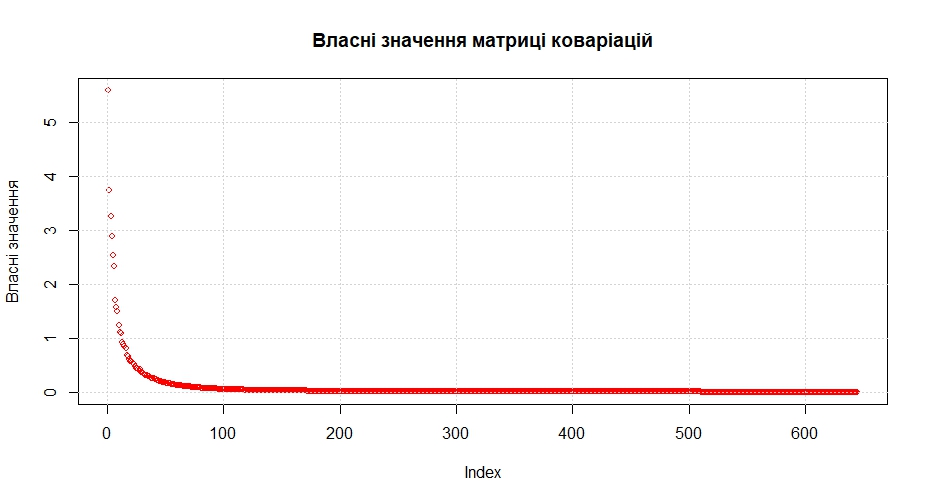
1. Для роботи з цією лабораторною роботою потрібно завантажити базу даних MNIST[[1]](#footnote-1). Дана база даних містить набір рукописних цифр від до . Перші 9 цифр ми можемо бачити на наступному графіку:

Даними в даній базі є бітні матриці, що зображають число (1 –білий колір на малюнку, 0 – сірий колір). Основним завданням п.1 є вивід 9 цифр з бази даних MNIST аналогічну малюнку 1.

1. Використовуючи базу даних MNIST побудувати коваріаційну матрицю для даних та вивести на графіку власні значення матриці . Для 55000 значень власні значення мають наступний вигляд:



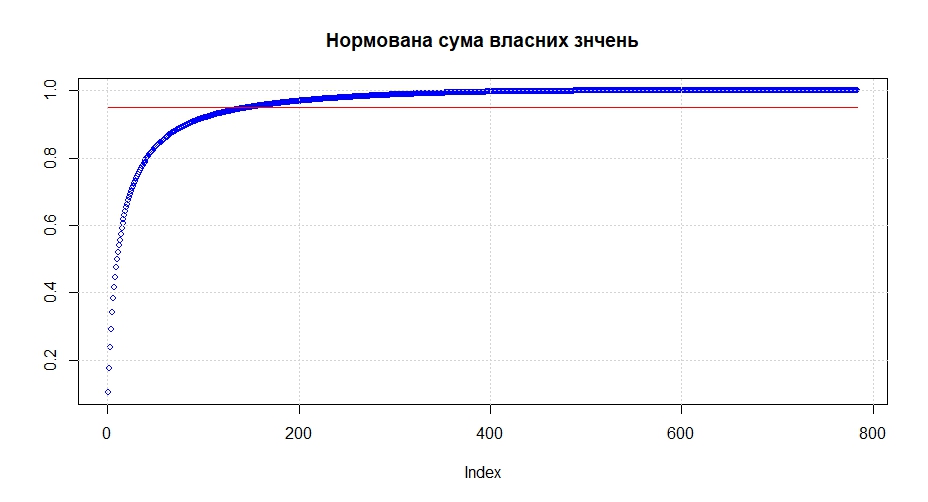
Як ми бачимо з даного малюнка, більшість власних значень рівна 0. Це пов’язано з тим, що краї малюнка завжди не використовуються для запису цифри, тобто в матриці ці значення завжди є 0. Тому з бази даних варто видалити всі пікселі, які завжди є 0, тобто видалити всі пікселі, дисперсія яких є 0. Кількість таких значень повинна бути в 650, тобто на практиці не використовується лише пікселя, що знаходяться на краях. Знайти всі власні значення для матриці, що містить лише значущі пікселі (не всі 0). Результат буде наступний:



Знайти кількість пікселів, які мають нульову дисперсію.

1. Використовуючи власні значення з п.3, знайти кількість пікселів, які зберігають відсотків розкиду початкових даних. Дану кількість можна знайти, використовуючи формулу:

де – загальна кількість власних значень матриці . Дану кількість можна визначити з наступного малюнка, на якому зображено



З даного малюнка видно, що потрібна кількість головних компонент знаходиться в межах 150. Тобто кількість потрібних факторів для аналізу та кластеризації, складає не пікселів, а лише приблизно 150 ‘узагальнених пікселів’ (головних компонент). Таким чином, вдалося скоротити задачу приблизно в 5 разів. Знайти дану кількість головних компонент.

1. Використовуючи всі дані та лише головних компонент, визначених в п.3, просте кластеризацію методом – середніх для . Визначити матриці невідповідностей для двох випадків. Зробити висновок про якість кластеризації на основі даних матриць. Матриця невідповідності буде мати наступний вигляд:

Labels 0 1 2 3 4 5 6 7 8 9

0 389 0 1 28 5 20 47 0 9 23

1 0 510 1 1 3 45 0 2 2 0

2 5 85 329 24 8 30 8 2 10 2

3 4 35 13 339 19 24 7 8 89 1

4 0 7 1 0 207 107 6 106 0 11

5 4 7 0 110 29 181 11 9 76 2

6 6 20 1 2 1 21 295 0 0 167

7 0 27 1 0 117 38 0 336 1 0

8 4 41 4 79 14 67 3 8 254 8

9 4 8 0 4 214 30 1 217 2 3

1. <https://ru.wikipedia.org/wiki/MNIST_(база_данных)>

   <http://yann.lecun.com/exdb/mnist/> [↑](#footnote-ref-1)